MonALISA, An Agent-Based Monitoring and Control System for the LHC Experiments

MonALISA

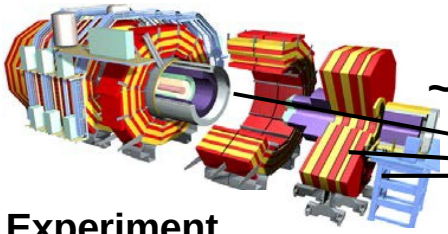
MONitoring Agents using a Large Integrated Services Architecture

Dorian Kcira

California Institute of Technology

CHEP, 22nd International Conference on Computing in High Energy Physics
San Francisco, October 2016

Early LHC Computing Model



Experiment

~PByte/sec

Online System

~500-1500 MBytes/sec

CERN Center PBs of Disk; Tape Robot

Tier 0 +1

Tier 1

10 – 40 to 100 Gbps

Tier1 Tier1 Tier1 Tier1 ...

Tier 2

Tier2 Center Tier2 Center Tier2 Center Tier2 Center Tier2 Center

Tier 3

10 to N X 10 Gbps

Institute Institute Institute Institute

Physics data cache

1 to 10 Gbps

Tier 4

Workstations

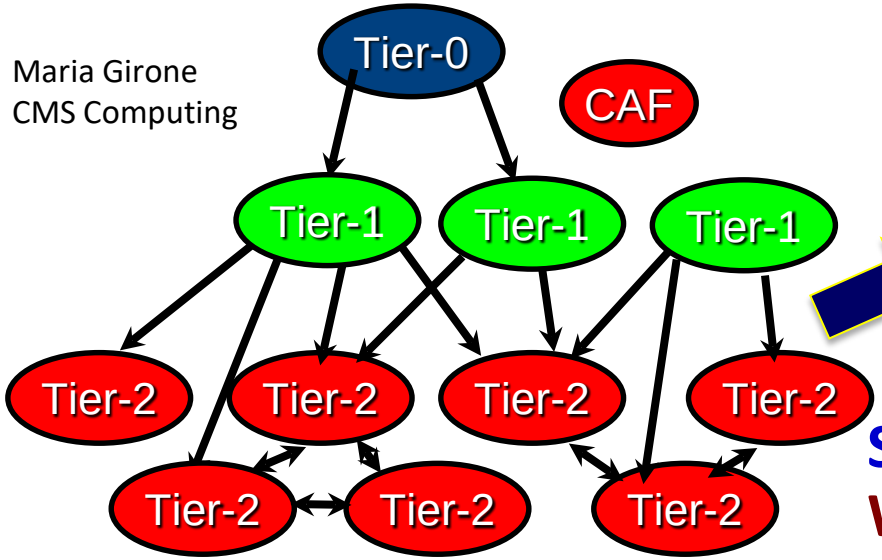
11 Tier1 and *160 Tier2 and Tier3 Centers*

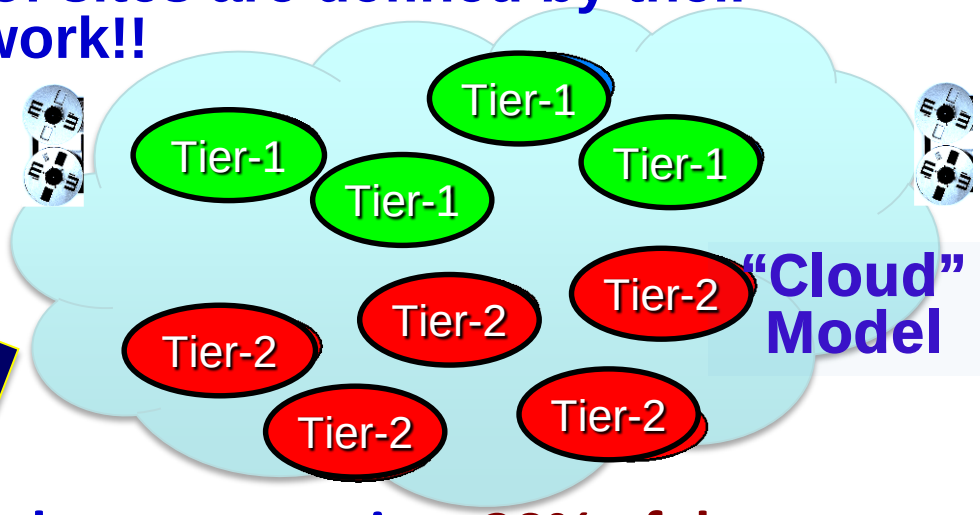
Tens of Petabytes by year for each experiment

Location Independent Access: Blurring the Boundaries Among Sites + Analysis vs Computing

- Once the archival functions are separated from the Tier-1 sites, the functional difference between Tier-1 and Tier-2 sites becomes small [and the analysis/computing-ops boundary blurs]
- Connections and functions of sites are defined by their capability, including the network!!

Maria Girone
CMS Computing





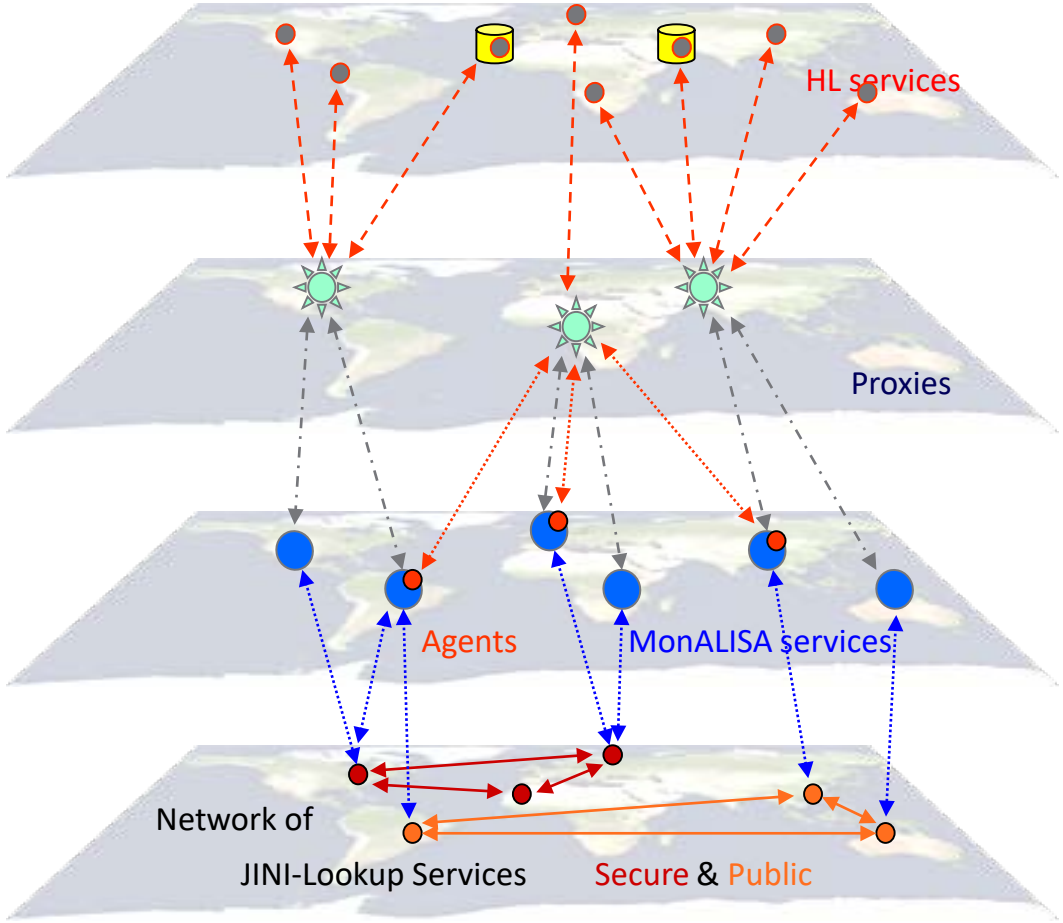
Scale tests ongoing: 20% of data across WAN: 200k jobs, 60k files, (100TB)/day

+Elastic Cloud-like access from some Tier1/Tier2/Tier3 sites

Monitoring Distributed Systems

- **MonALISA: Monitoring Agents in A Large Integrated Services Architecture**
- **An essential part of managing large scale, distributed data processing facilities, is a monitoring system that is able to monitor computing facilities, storage systems, networks and a very large number of applications running on these systems in near-real time.**
- **The monitoring information gathered for all the subsystems is essential for design, modelling, debugging, accounting and the development of higher level services, that provide decision support and some degree of automated decisions and for maintaining and optimizing workflows in large scale distributed systems.**

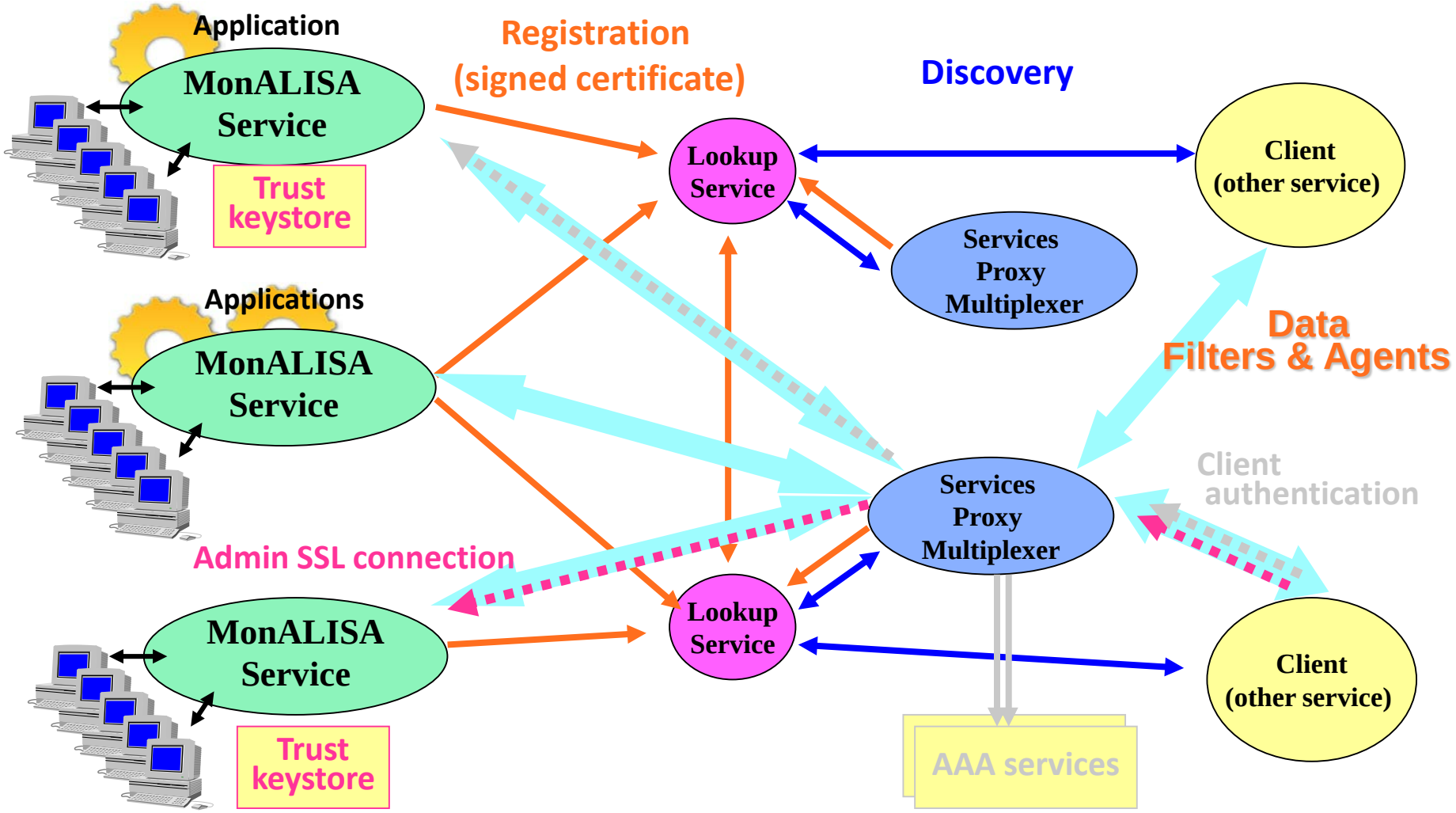
The MonALISA Architecture



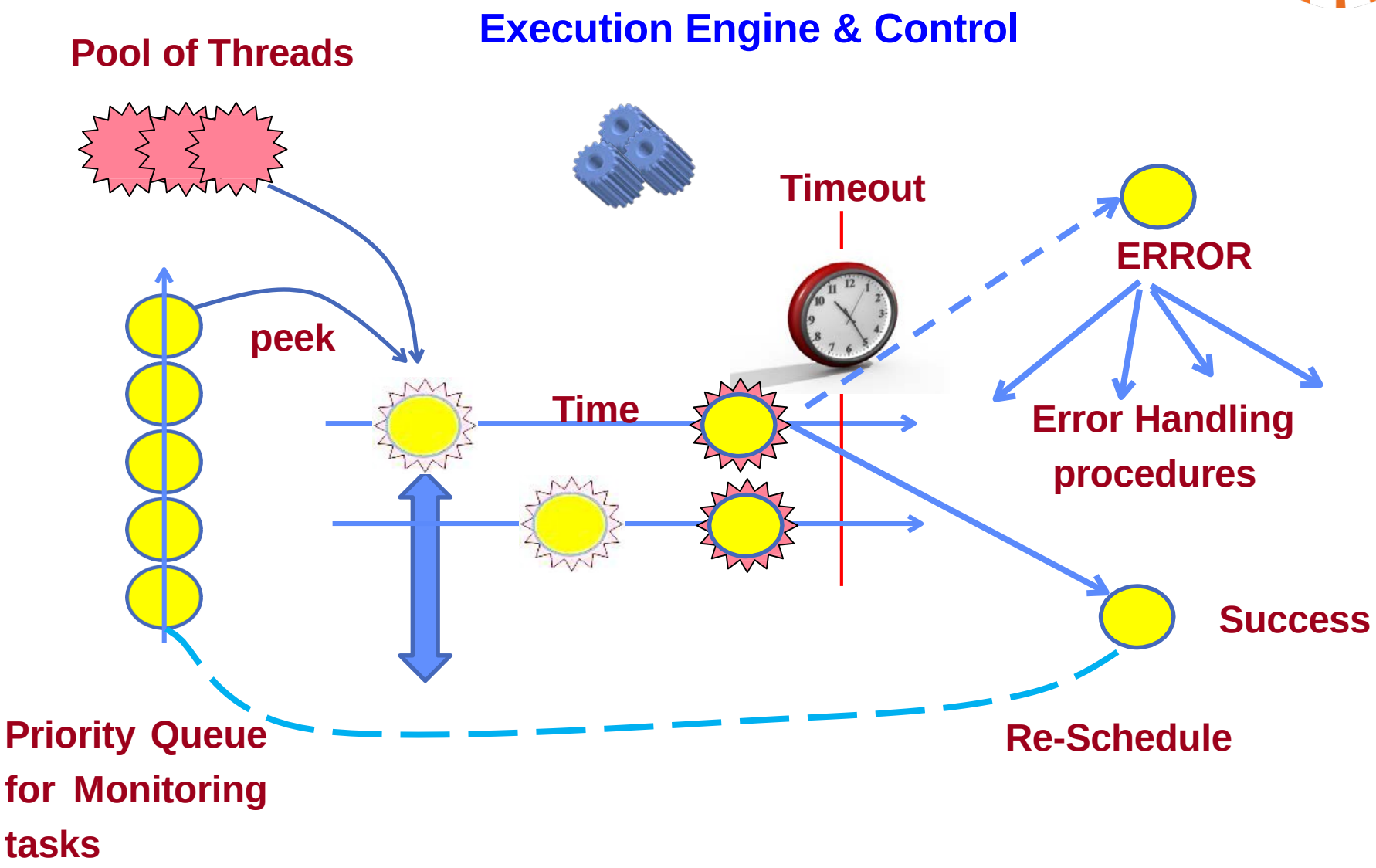
Fully Distributed System with no Single Point of Failure

- **Regional or Global High Level Services, Repositories & Clients**
- **Secure and reliable communication**
- **Dynamic load balancing**
- **Scalability & Replication**
- **AAA for Clients**
- **Distributed System for gathering and analyzing information based on mobile agents: Customized aggregation, Triggers, Actions**
- **Distributed Dynamic Registration and Discovery-based on a lease mechanism and remote events**

Registration / Discovery
Admin Access and AAA for Clients

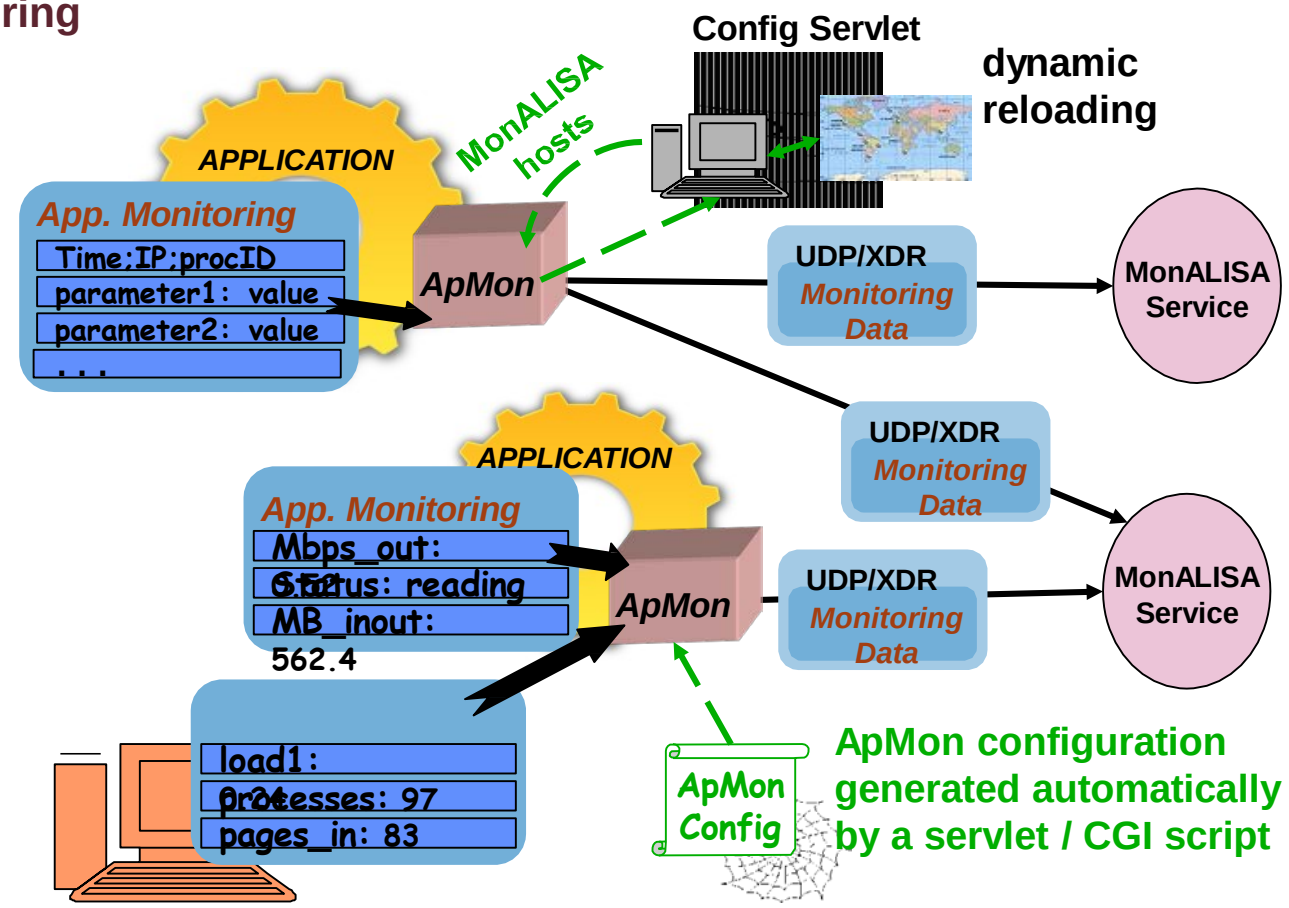


Multi-thread Execution Engine



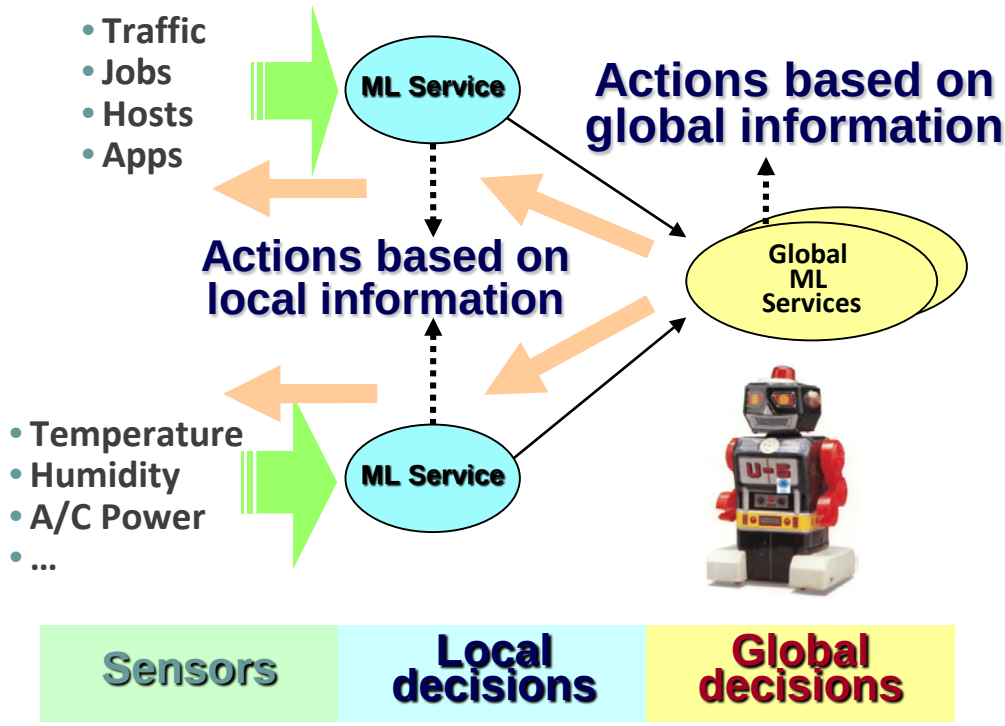
ApMon – Application Monitoring

- UDP based Library of APIs (C, C++, Java, Perl. Python) that can be used to send any information defined by users or applications to MonALISA services
- Flexibility, dynamic configuration, high communication performance
  - Automated system monitoring
  - Accounting information



Local and Global Decision Framework

- Two levels of decisions:
  - local (autonomous)
  - global (correlations)
- Actions triggered by:
  - values above or below given thresholds
  - Absence or presence of values
  - correlations between any values
- Action types:
  - alerts (emails, instant messages, feeds)
  - automatic charts annotations in the repository
  - running custom code, like securely ordering MLs service to change connectivity – optimize traffic, submit jobs, (re)start global service

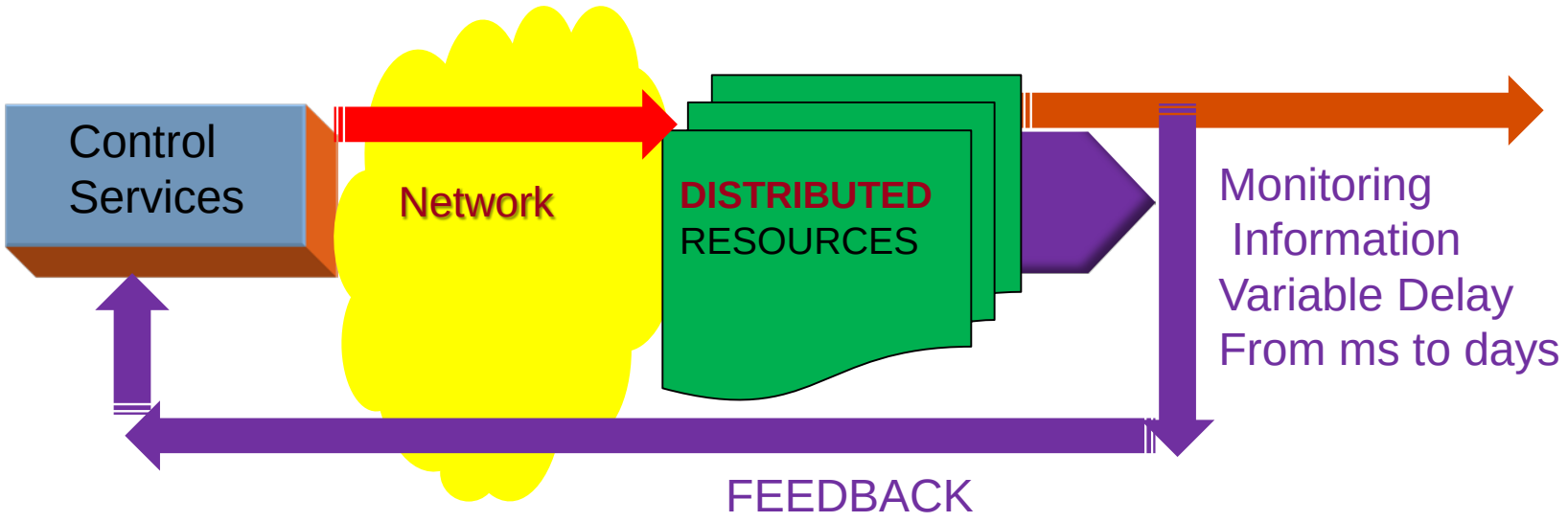


Control and Optimization

Time delays in receiving monitoring data for the control units :

- **give rise to phase lag**
- **degenerate system stability and performance**

Maximize temporal determinism. In general, a time-lag in a feedback loop will result in overshoot and oscillation. These oscillation could fade out, continue or increase to bring the system into an unstable state.



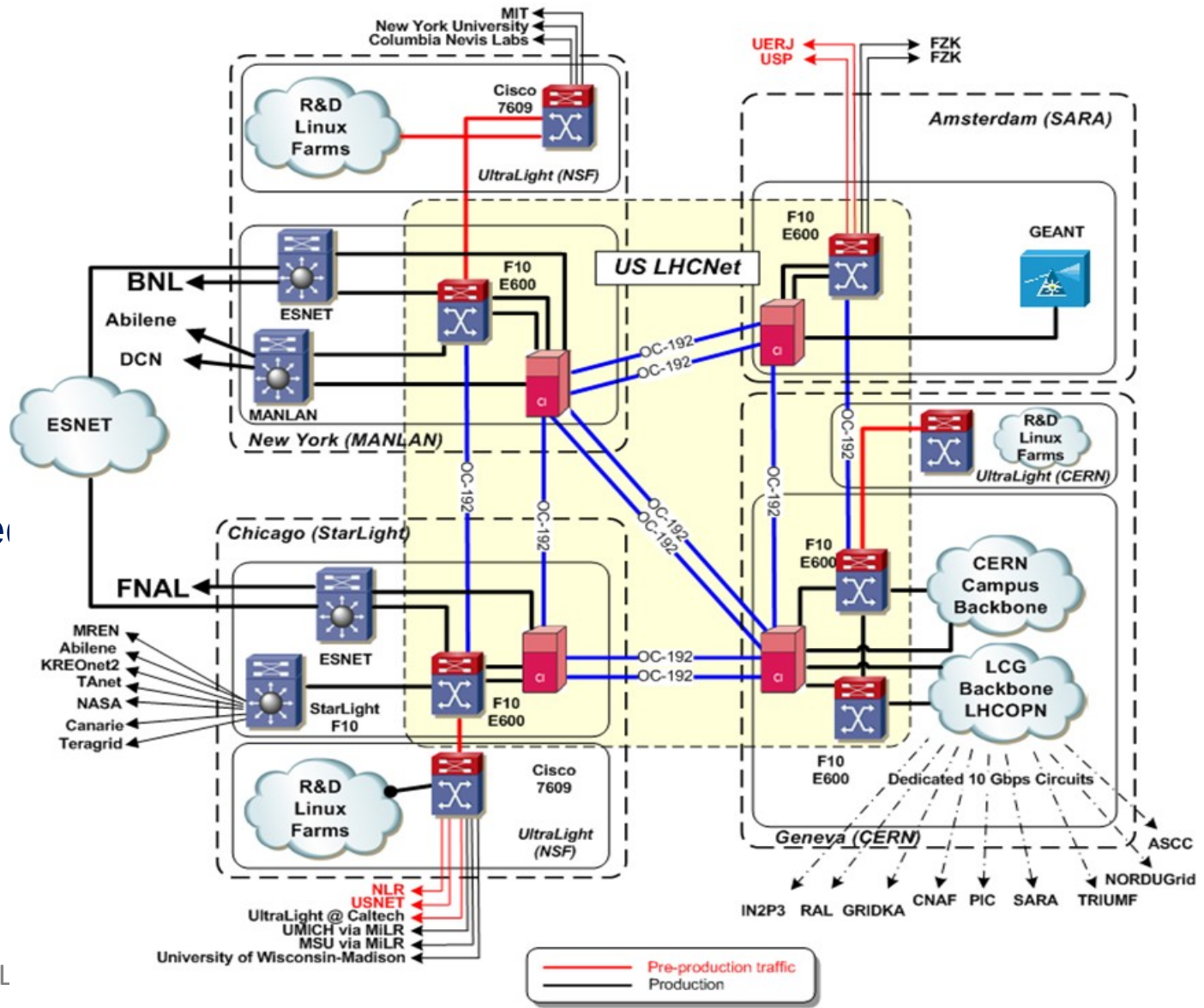
Package & Information Collected

- The MonALISA package includes:
  - Local host monitoring (CPU, memory, network traffic , processes and sockets in each state, LM sensors, APC UPSs), log files tailing
  - SNMP generic & specific modules
  - Condor, PBS, LSF and SGE (accounting & host monitoring), Ganglia
  - Ping, tracepath, traceroute, pathload and other network-related measurements
  - TL1, Network devices, Ciena, Optical switches
  - Calling external applications/scripts that return as output the values
  - XDR-formatted UDP messages (ApMon – user's defined information).
- New modules can be easily added by implementing a simple Java interface. Filters can be used to generate new aggregate data.
- The Service can also react to the monitoring data it receives (actions & alarms).
- MonALISA can run code as distributed agents for global optimization

Monalisa Monitoring
Networks: USLHCNet

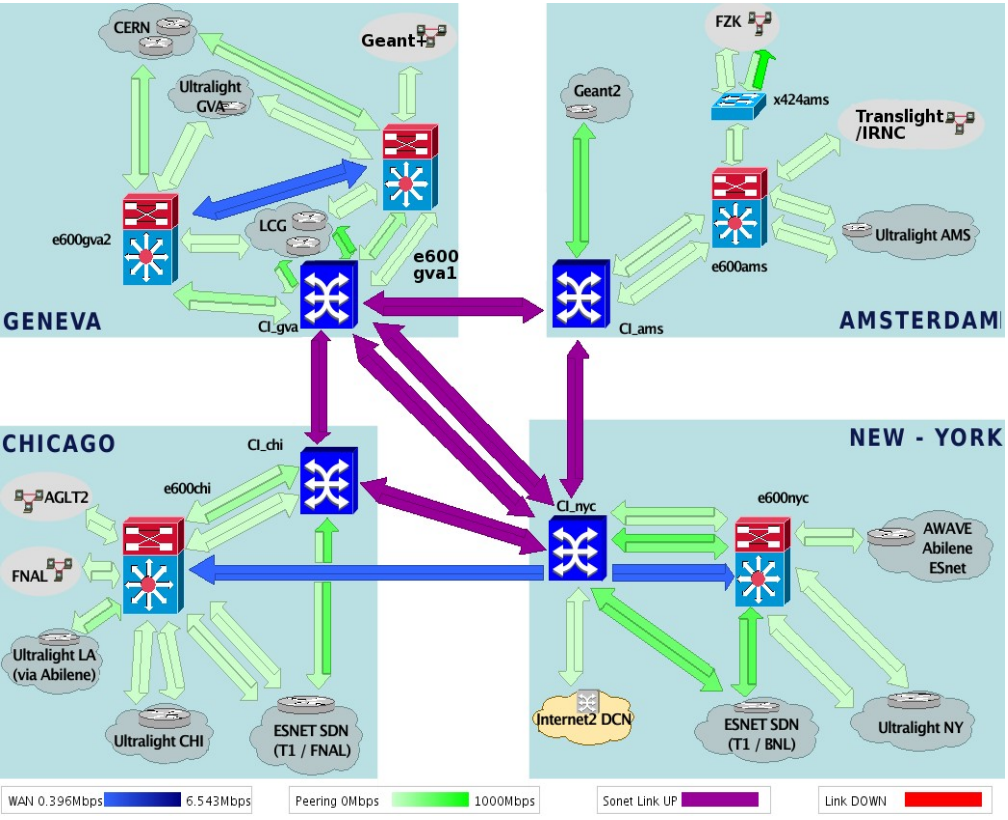
USLHCNet: High-Speed Trans-Atlantic Network

- CERN to US
  - FNAL
  - BNL
- 6 x 10G links
- 4 PoPs
  - Geneva
  - Amsterdam
  - Chicago
  - New York
- The core is base on Ciena CD/CI (Layer 1.5)
- Virtual Circuits

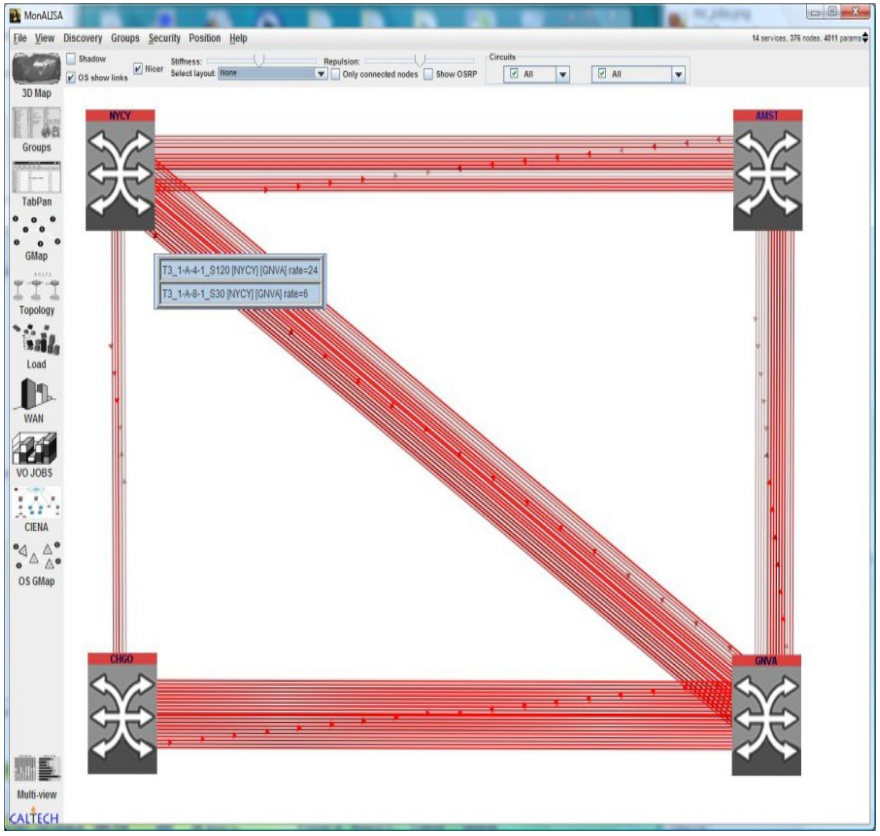


USLHCNet: Monitoring the Topology

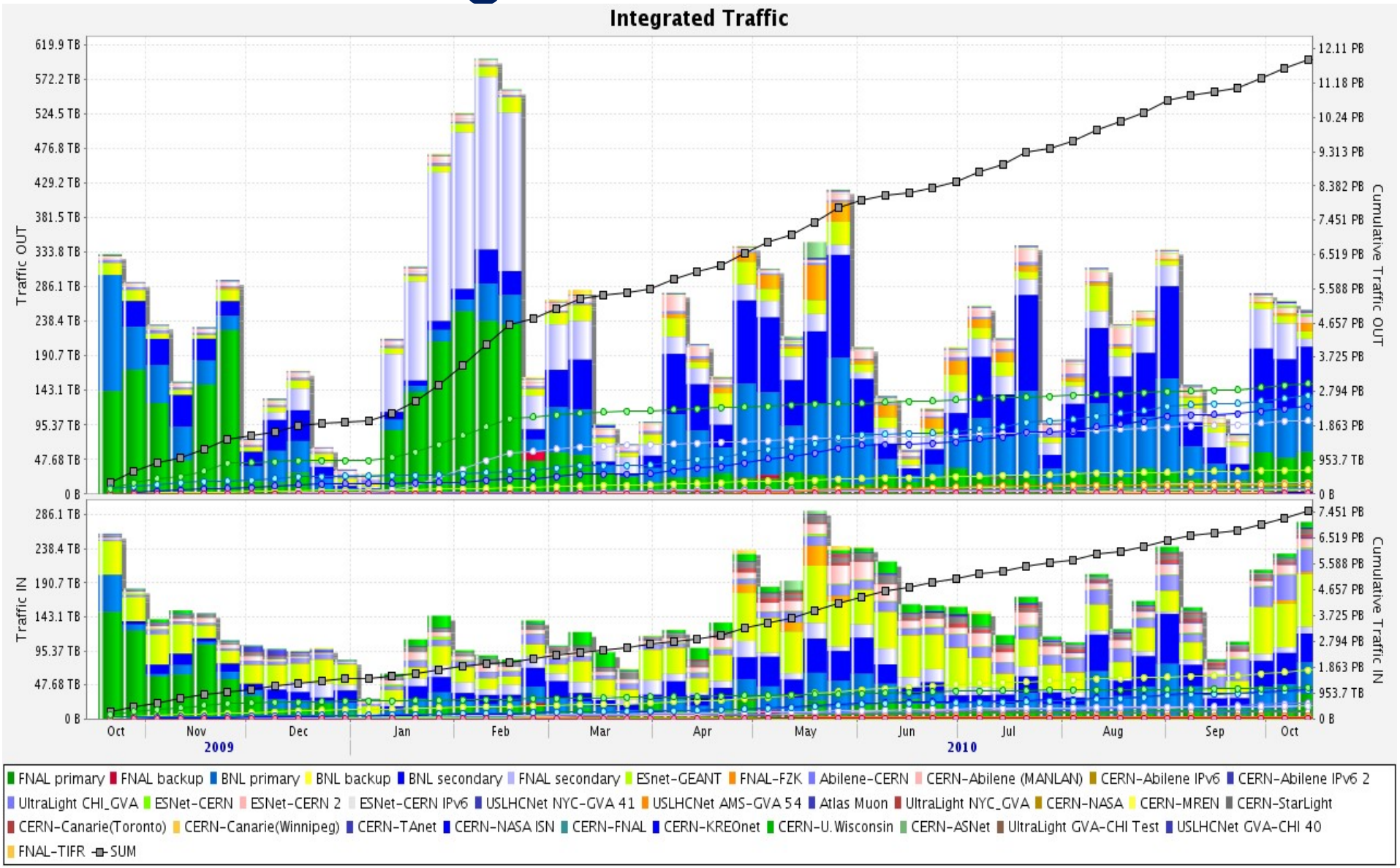
Topology, Status, Peering



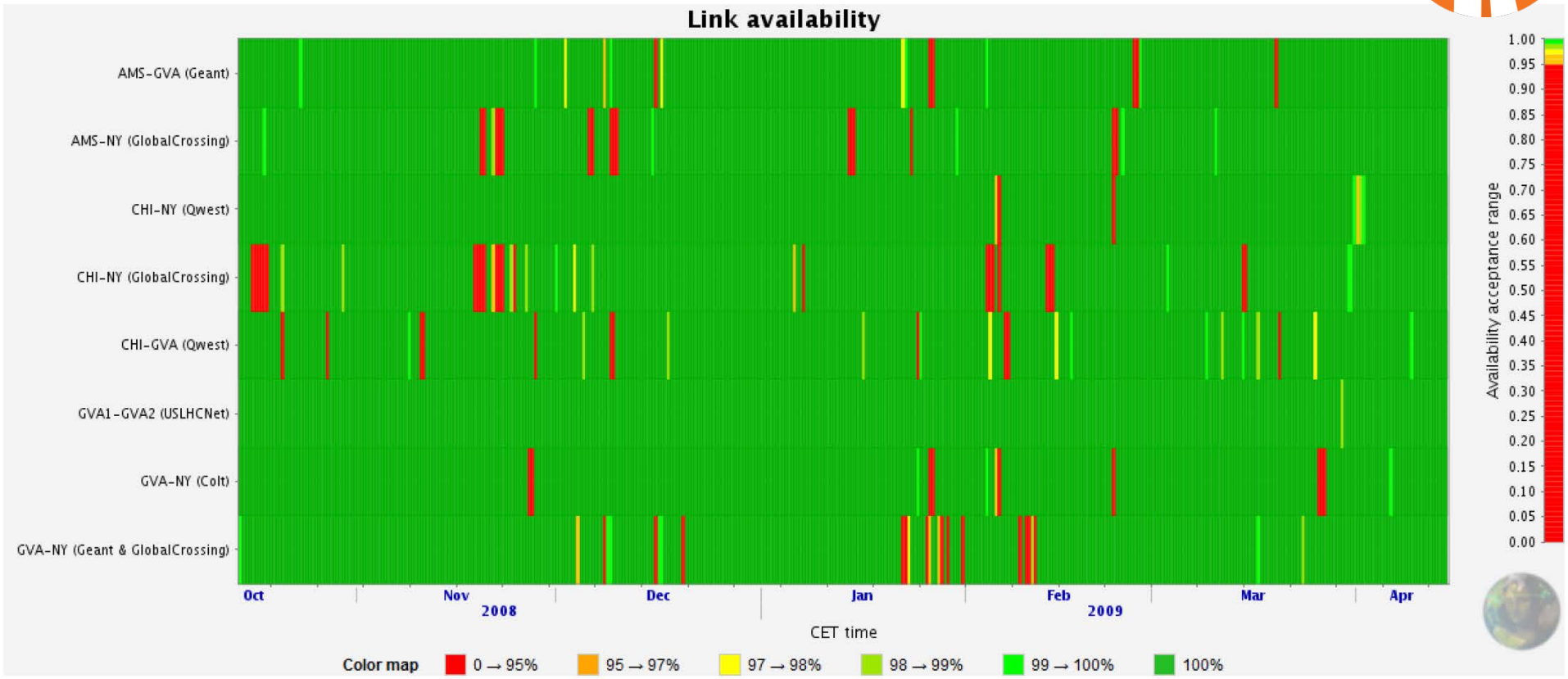
Real Time Topology for L2 Circuits



USLHCNet: Accounting for Integrated Traffic



USLHCNet: Link Availability

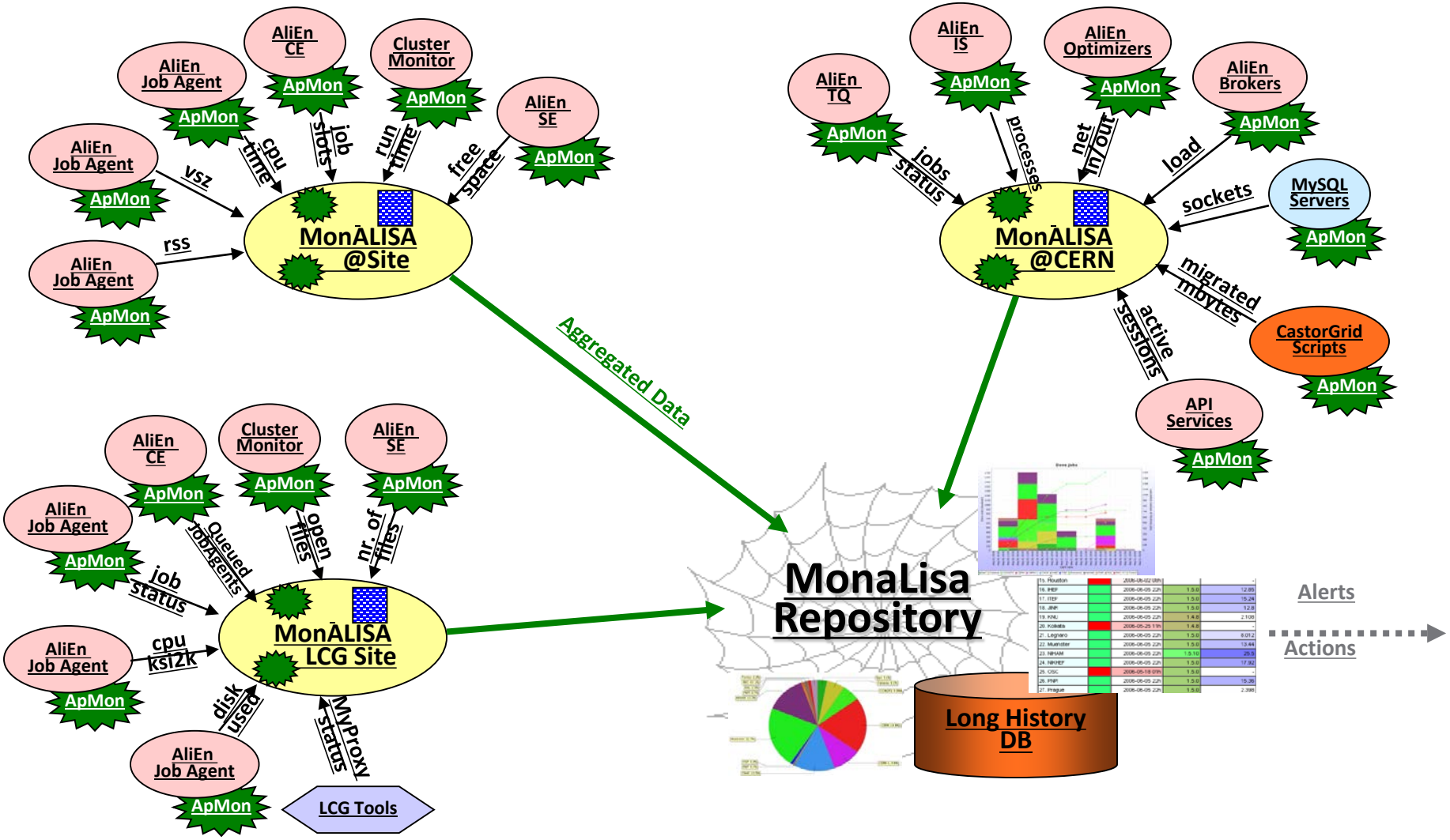


Statistics

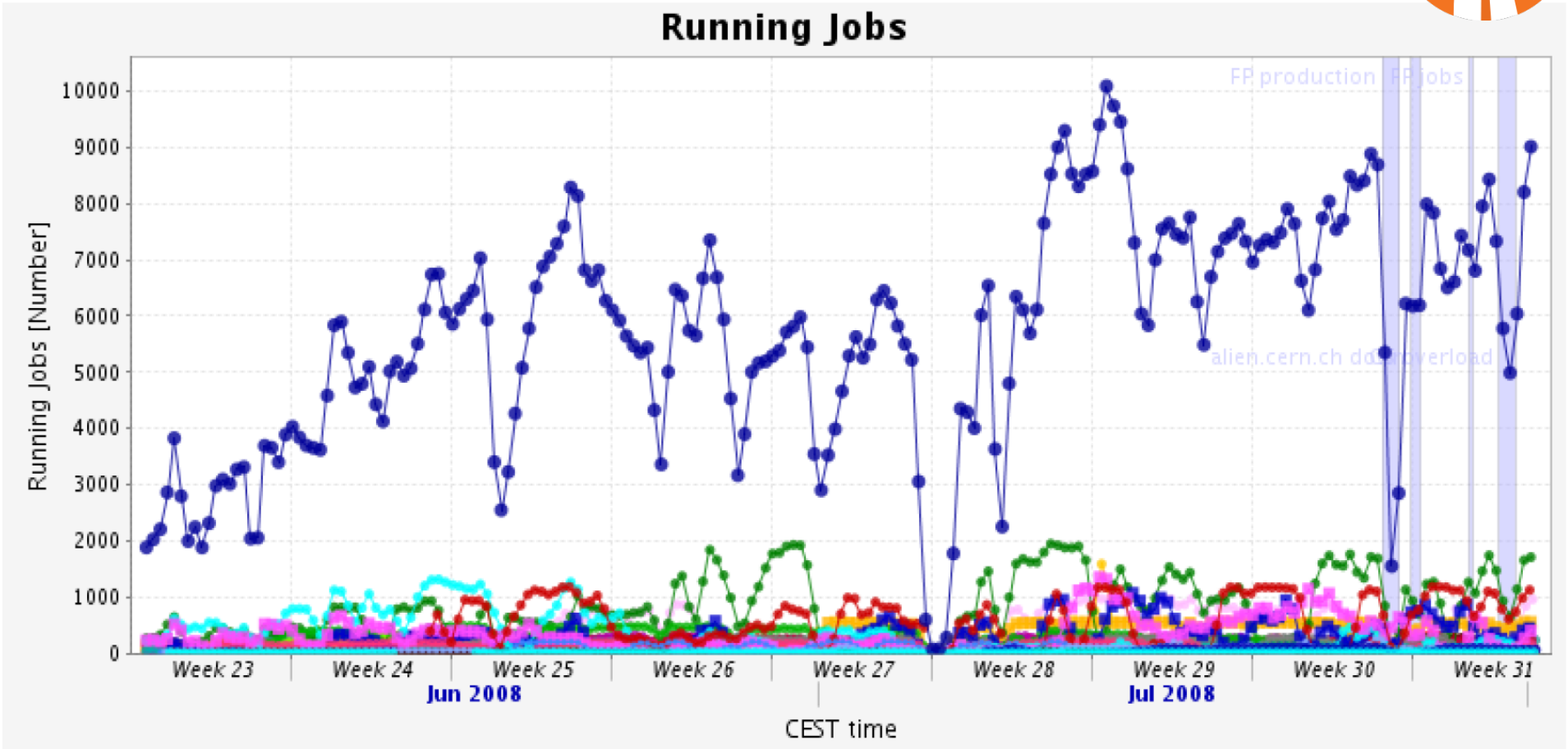
Link name	Data		Monitoring		Link
	Starts	Ends	Availability(%)	Gaps	Availability(%)
AMS-GVA (Geant)	14 Oct 2008 12:22	14 Apr 2009 12:21	99.100%	4m 30s	99.53%
AMS-NY (GlobalCrossing)	14 Oct 2008 12:22	14 Apr 2009 12:21	100%	-	97.87%
CHI-NY (Qwest)	14 Oct 2008 12:22	14 Apr 2009 12:21	99.93%	2:59	99.90%
CHI-NY (GlobalCrossing)	14 Oct 2008 12:22	14 Apr 2009 12:21	99.62%	16:40	96.59%
CHI-GVA (Qwest)	14 Oct 2008 12:22	14 Apr 2009 12:21	99.100%	4m 31s	99.29%
GVA1-GVA2 (USLHCNet)	14 Oct 2008 12:22	14 Apr 2009 12:21	100%	-	99.100%
GVA-NY (Colt)	14 Oct 2008 12:22	14 Apr 2009 12:21	100%	-	98.91%
GVA-NY (Geant & GlobalCrossing)	14 Oct 2008 12:22	14 Apr 2009 12:21	99.99%	13m 28s	99.47%

Monalisa Monitoring
ALICE Distributed Computing Environment

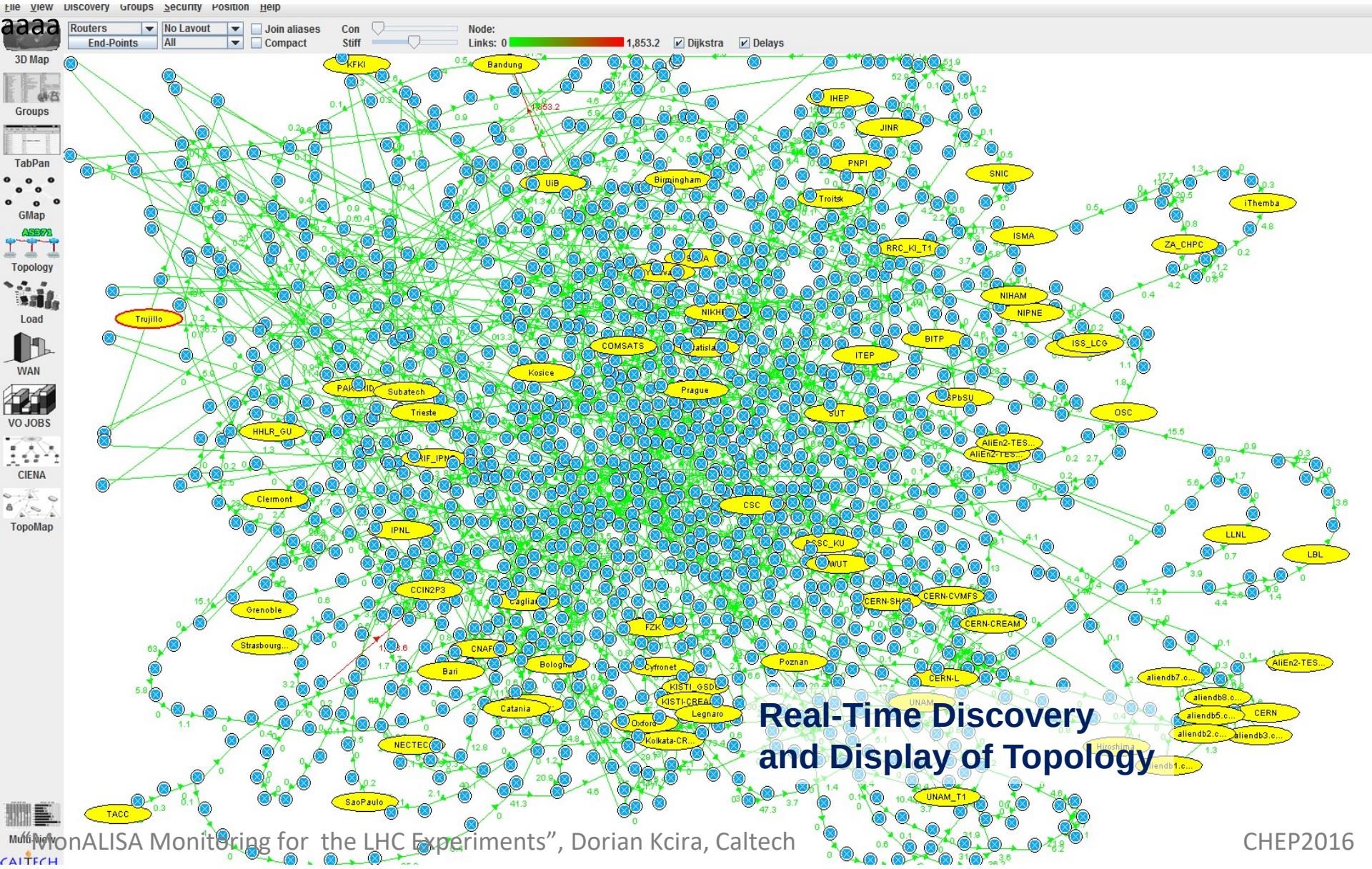
Monitoring in ALICE



ALICE Running Jobs



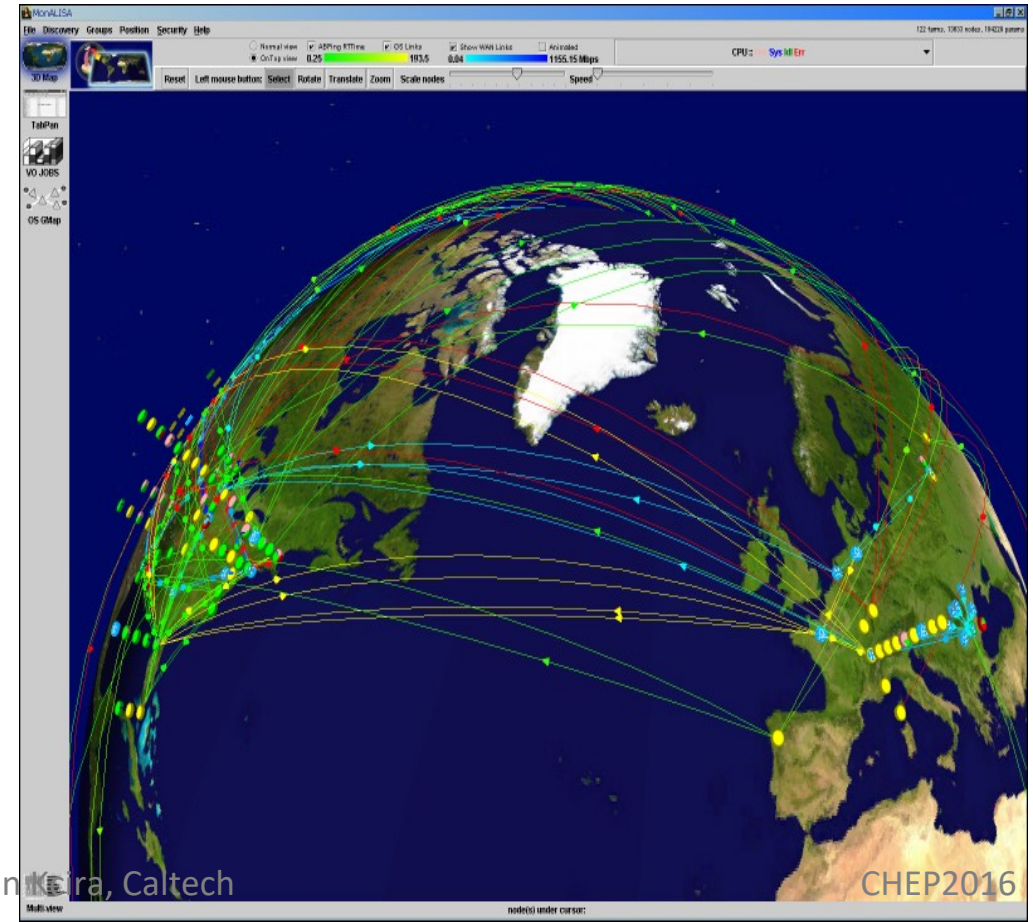
ALICE: Network Topology (L3), Latency & Routers for all Sites



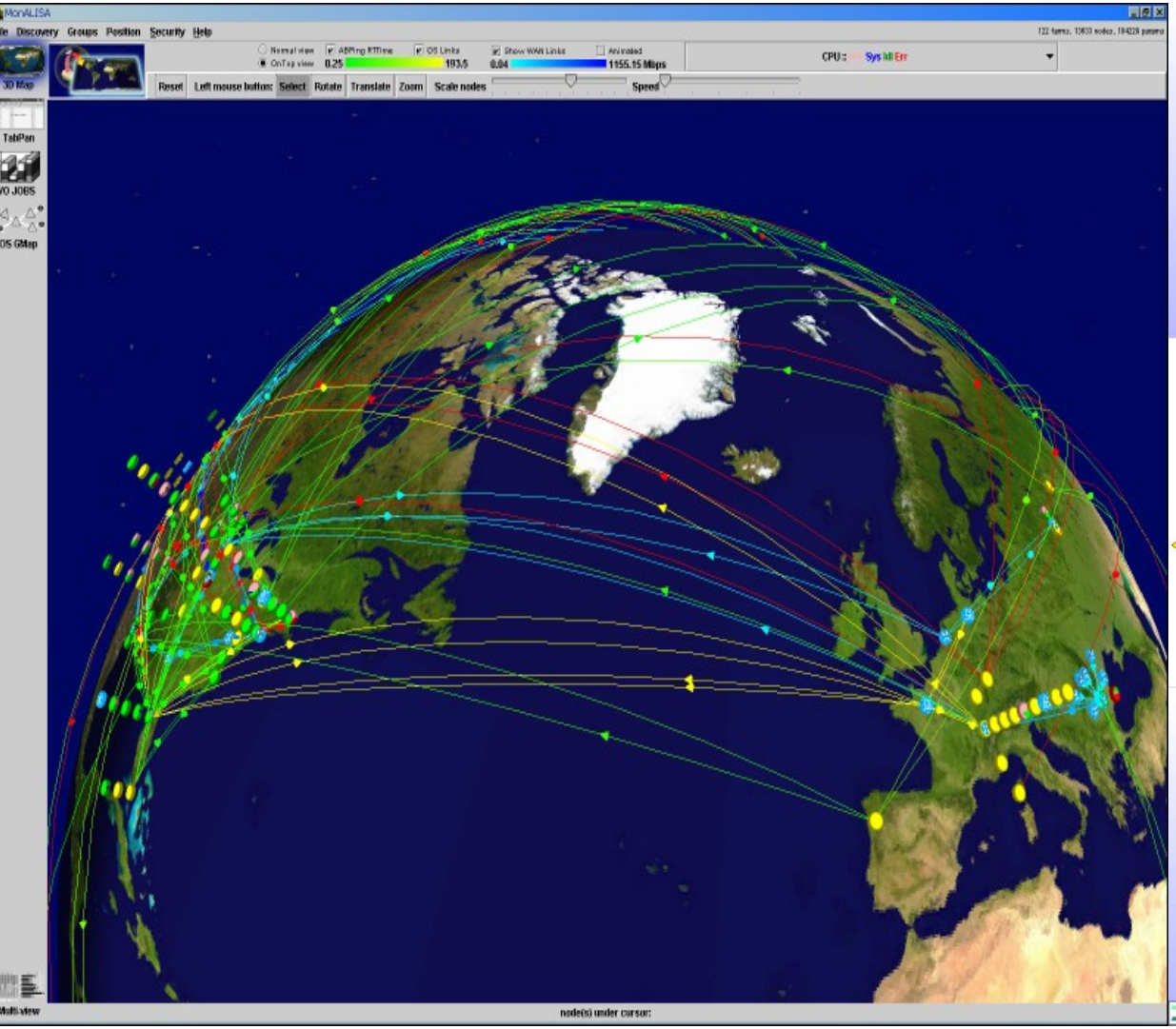
MonALISA Monitoring Today

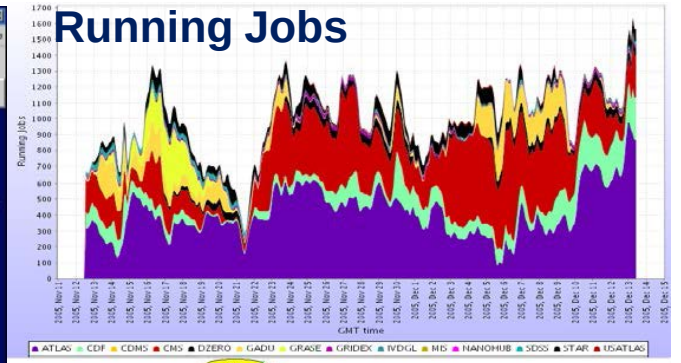
- 60K computers
- > 100 Links of Major Netws
- Tens of Thousands of Grid jobs running concurrently
- 14 K end-to-end network path measurements
- Using Intelligent Agents
- Collecting 6 million persistent parameters in real-time
- 100 millions of volatile parameters per day
- Updating 35K parameters per second
- Repository servers 10M users request / year

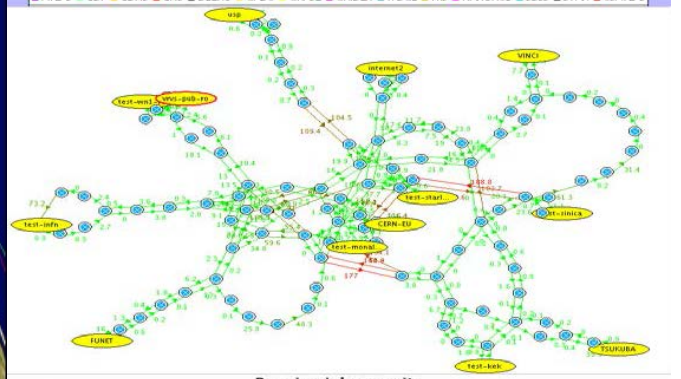
Running 24x7 at more than 370 sites

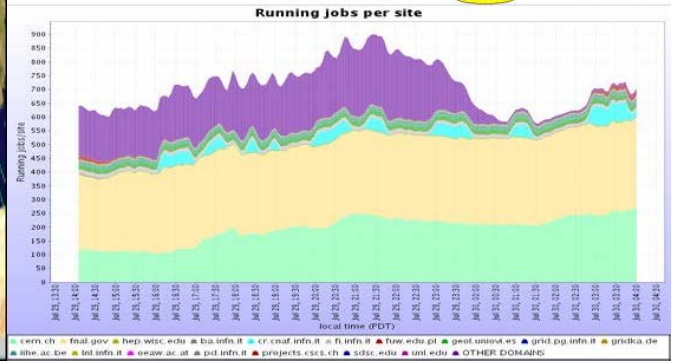


Grid Sites, Running Jobs, Network Traffic, Connectivity









MonALISA Summary

- MonALISA provides a unified platform for monitoring information for local and distributed systems
- Service Oriented Architecture & Dynamic Discovery
- Agent model for monitoring modules, filters and actions
- Dynamic, on-the-fly subscription to services and information sources
- Simple and efficient communication approach (problems with RMI, XML, etc)
- Multithreading instrumental for performance and reliability of the system
- Various graphical views to displays to present information
- Simple and efficient approach for storing data

MonALISA, Further Info

- ***Grid and Cloud Computing: Concepts and Practical Applications*, Carminati, F., Betev, L., Grigoras, A., ISBN 978-1-61499-642-2**
- **International School of Physics "Enrico Fermi", Varenna, 2014**
  - **http://static.sif.it/SIF/resources/public/files/va2014/Legrand.pdf**
- **CHEP 2016 Presentation:**
  - ***Review of Terabit/sec SDN demonstrations at Supercomputing 2015 and Plans for SC16*, Azher Mughal, Id 271**
- **CHEP 2016 Posters:**
  - ***SDN-NGenIA A Software Defined Next Generation integrated Architecture for HEP and Data Intensive Science*, D. Kcira, Id 79**
  - ***Data Transfer Nodes and Demonstration of 100G - 400G Wide Area Throughput Using the Caltech SDN Testbed*, Azher Mughal, Id 272**